Metadata of the chapter that will be visualized in SpringerLink

Book Title	Digital Transformation and Global Society	
Series Title		
Chapter Title	A Framework for Intelligent Policy Decision Making Based on a Government Data Hub	
Copyright Year	2020	
Copyright HolderName	Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Al-Lawati**
	Particle	
	Given Name	**Ali**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Information Technology Authority
	Address	Muscat, Oman
	Email	ali.allawati@ita.gov.om
Author	Family Name	**Barbosa**
	Particle	
	Given Name	**Luis**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	United Nations University Operating Unit on Policy-Driven Electronic Governance
	Address	Guimarães, Portugal
	Email	barbosa@unu.edu

Abstract	The e-Oman Integration Platform is a data hub that enables data exchanges across government in response to transactions. With millions of transactions weekly, and thereby data exchanges, we propose to investigate the potential of gathering intelligence from these linked sources to help government officials make more informed decisions. A key feature of this data is the richness and accuracy, which increases the value of the learning outcome when augmented by other big and open data sources. We consider a high-level framework within a government context, taking into account issues related to the definition of public policies, data privacy, and the potential benefits to society. A preliminary, qualitative validation of the framework in the context of e-Oman is presented. This paper lays out foundational work into an ongoing research to implement government decision-making based on big data.

Keywords	Government big data - Policy making - Data analysis - e-Oman

A Framework for Intelligent Policy Decision Making Based on a Government Data Hub

Ali Al-Lawati[1(✉)] and Luis Barbosa[2]

[1] Information Technology Authority, Muscat, Oman
ali.allawati@ita.gov.om
[2] United Nations University Operating Unit on Policy-Driven Electronic Governance, Guimarães, Portugal
barbosa@unu.edu

Abstract. The e-Oman Integration Platform is a data hub that enables data exchanges across government in response to transactions. With millions of transactions weekly, and thereby data exchanges, we propose to investigate the potential of gathering intelligence from these linked sources to help government officials make more informed decisions. A key feature of this data is the richness and accuracy, which increases the value of the learning outcome when augmented by other big and open data sources. We consider a high-level framework within a government context, taking into account issues related to the definition of public policies, data privacy, and the potential benefits to society. A preliminary, qualitative validation of the framework in the context of e-Oman is presented. This paper lays out foundational work into an ongoing research to implement government decision-making based on big data.

AQ1

Keywords: Government big data · Policy making · Data analysis · e-Oman

1 Introduction

1.1 Aims and Motivation

The increase in the amount of data generated in different forms and across different domains has presented new opportunities to make inferences that go beyond typical e-commerce applications. Recently, there has been an increasing interest in using big data to make inferences on social trends for policy decision making.

However, a key obstacle to the usage of big data for this purpose is the haphazard methods by which big data is often created. Big data in most cases is not collected in a scientifically sound process. As such, it may not be representative of the population of interest and using it without sufficient care can result in misguided inferences. The infamous case of Google Flu Trends in 2010 where the algorithm overestimated the outbreak of the flu by 100% [1] is an often cited example.

When big data is combined with linked data that is semantically query-able, the added-value can be significantly improved [2]. The e-Oman Integration Platform is a data hub that currently supports the exchange of data across many government organizations

D. A. Alexandrov et al. (Eds.): DTGS 2019, CCIS 1038, pp. 1–15, 2020.
https://doi.org/10.1007/978-3-030-37858-5_8

in real-time. In this paper, we investigate whether this data, when combined with big, open, and social data can become a valuable tool for policy-making [3].

The purpose of this paper is to analyze the challenges and considerations of performing this task by conducting a systematic literature review on big data usage for policy-making in government. Based on our findings, we present a framework for the design of data-driven applications within a governmental context.

We discuss the benefits of such an approach, as well as constraints related to privacy concerns and efficacy to the definition of public policies. The framework was motivated by the concrete need to foster the productive use of data managed by the e-Oman platform. A preliminary validation was done resorting to two key stakeholders.

This paper is structured as follows. The rest of this section describes the e-Oman Integration platform and defines the problem statement within the context of e-Oman. Section 2 details the methodology for literature review and describes an automated tool implemented to support intelligent clustering over a base of articles. A literature review follows in Sect. 3 based on the relevant themes identified. Section 4 describes the four-step framework, which is analytically validated in Sect. 5 by remote interviews with two key stakeholders of the e-Oman Integration Platform. Finally, Sect. 6 concludes and discusses future work.

1.2 Context

e-Oman is an umbrella project that aims to advance Oman into a digital society. It encompasses the enhancement of government services from traditional paper-based and manual processes into electronic and online services. A key obstacle is the exchange of data relevant to the service across different government organizations. This has been addressed with the introduction of the Integration Platform.

The e-Oman Integration Platform is a data hub through which millions of data exchanges occur every week in real-time between different government agencies. It enables the exchange of current data across government while avoiding the burden of data duplication. Since data sets may be owned by different government entities, it has enabled inter-organizational services and burden reduction on residents and businesses alike. Data exchanges within the Integration Platform occur in a response to a transaction occurring at a government entity that requires additional information from a second one. There are currently more than 12 data providers and over 35 data consumers.

1.3 Problem Statement

The vast number of data exchanges that occur on a daily basis presents many opportunities to gather intelligence that would help government officials in the decision-making process. The paper discusses how this can be utilized in practice and which sort of framework may guide the design of such data-driven applications, eventually combining different data sources.

We believe an added value to society, potentially to promote sustainable development, can arise from the smart use of data that is already being collected by

governmental agencies or services, as in the case of Oman. This can be most valuable, for example, in the monitoring of population shifts, demographic changes, or development movements. Data in turn can be used by government officials to adjust budget allocations or aid in the planning process at local or national levels.

1.4 Scope

This paper is part of an ongoing research on the implementation of big data analytics for government decision making. While the problem statement spans the ultimate outcome of this research, this paper is concerned with the research work that lays the foundation for the ongoing research.

2 Methodology

2.1 Overview

Our research is sustained by a literature review that is based on systematic literature review standards. A specific approach to support paper categorization was devised for this work and implemented as an intelligent automated learning model. This system and its prototype implementation are further detailed in Subsect. 2.2.

A systematic literature review proceeds by specifying a "criterion-based selection" [4]. This establishes an objective selection method free from the biases that may emerge if the criteria are not clearly specified. In order to adhere to this definition of systematic literature review, we performed a search on "Scopus" for Journal articles where the terms "government" and "big data" appear in the title, abstract, or keywords. The choice of using "Scopus", self-proclaimed as "the largest abstract and citation database of peer-reviewed literature" [5] was motivated by the fact that Scopus has higher coverage of recent Journal articles [6]. While literature reviews often resort to keyword searches on multiple databases, restricting our search to a single database limits duplication, which might affect the performance of our learning model prototype.

Our search was limited to literature that specifically addresses the design of data-driven applications in the context of government, given our problem is specifically related to the usage of big data for the purposes of policy definition, planning, and decision making in a government context. While similar problems may have been addressed in the business sector, the mission, requirements, and challenges within a government context are different from the business sector. For example, a key challenge quite specific to this context involves collecting data from various and often competing agencies [7] all in the public sector.

Formally, only peer-reviewed journal articles were considered, which is a common standard as in [4, 8], and [9], and limits our result set to 669 articles. However, following the guidelines in [10], our evaluation was not confined to a small sample of top journals. Instead, we included all published articles that satisfied our search terms in a preliminary evaluation regardless of the field and the prominence of the journal.

Once we completed the literature review, we used the findings to define the envisaged framework for designing governmental, data-driven applications.

The last stage consisted of a preliminary validation through exposition and discussion to two senior officers from the Information Technology Authority of Oman. Although a more systematic test of the framework, through a pilot application to a concrete case study is planned, this qualitative verification provided an early, informed feedback from the real, final stakeholders, and guide subsequent developments.

2.2 The Automated Learning Model

We develop a specific method for automated support to paper evaluation in the form of a smart algorithm. This is based on a learning model that classifies each paper into a cluster of articles under a similar general theme. The system helps to summarize the literature into clusters, and structures the content, such that the most relevant are not missed. As a result, more attention is dedicated to articles most germane to the objectives fixed for this research; while objectively including literature from different domains, which is most appropriate when an interdisciplinary field such as big data is being addressed. Specifically, the purpose of the learning model is to identify the main general themes in the literature where the keys *government* and *big data* are combined. However, we realize that most such articles do span more than one theme. While we structure the literature review based on the outcome of the learning model, we consider the topics in each article related to clusters other than the one it was associated to.

Clustering is a well-known class of algorithms in machine learning whereby a set of objects are grouped in a way such that similar objects are grouped together. Clustering can be further classified into supervised clustering, and unsupervised clustering. In supervised clustering, the target clusters are defined beforehand and supplied to the algorithm, whereas in unsupervised clustering, it is up to the algorithm to define the clusters based on a pre-defined number of clusters. In the interest of not presupposing the topics in the literature, we elect to perform unsupervised clustering using a well-known algorithm: k-means.

Prior to clustering, the set of articles (dataset) gathered needs to be quantified. One method for performing this is using a bag-of-words approach combined with a scoring function such as TFIDF. While this can be an effective method in some cases, it can often provide a negative outcome, particularly where the distinction between articles is not so concrete, as it happened in our case. The bag-of-words approach fails to take context into consideration; instead it merely considers keywords and gives more weights to rare keywords [11].

Thus, we opt for paragraph vectors proposed by Le and Mikilov [11], which learns "fixed-length feature representations from variable-length pieces of texts". We generate these vectors upon a concatenation of the title, keywords, and abstract of an article.

Scopus provided training data for the paragraph vectors model. We exported a set of 26938 documents based on the search term "big data". The set was fed to the "gensim doc2vec" library which is a well-tested implementation of paragraph vectors [12].

The duration of the training was 4 min 35 s, for our training set. The model was verified by comparing inferred vectors with the training corpus. As the authors of gensim suggest in their documentation: to assess the model one can use the "training corpus as some new unseen data and then seeing how the compare with the trained model" based on

self-similarity [13]. The model found that over 99.94% of documents are most similar to itself which suggests the model is "behaving in a consistent manner" [13].

The most similar and dissimilar documents for a random set of documents was analyzed and verified manually. For instance, given the article "Data Intelligence for Local Government? Assessing the Benefits and Barriers to Use of Big Data in the Public Sector", the most similar article to it based on this model is "New development: Leveraging 'big data' analytics in the public sector", and the least similar article is "A review on document image analysis techniques directly in the compressed domain".

The algorithm runs on the vector representation of the data set that meet our criterion-based selection. The elbow method is used to determine the right number of clusters most appropriate for the dataset (Fig. 1).

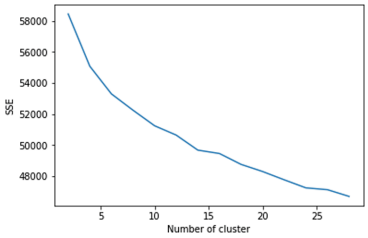

Fig. 1. The error sum of squares as a function of the number of clusters.

We selected a cluster of 10 based on the elbow method [] which establishes the optimal number of clusters. Table 1 presents the number of articles in each cluster along with a cluster name assigned to each cluster based on the prevailing topic contained articles.

Table 1. Clusters generated by the learning model and number of articles within each cluster.

Cluster	Papers
Sentiment analysis	31
High-level benefits	66
Data quality	91
Intelligent applications in the health sector	50
Big-data in government	58
Case studies of big-data in smart cities	59
Big government	106
Data privacy	72
Big data applications and strategy in the health sector	51
Policy and governance	54

Analysis was restricted to the clusters directly relevant to the objectives of this paper, although a cursory review of all other papers was also performed, including: high-level benefits, data quality, big government, big-data in government, data privacy, and policy and governance. We further reduced the list of papers based on an inspection by the authors to a set of 53 documents that is a representative sample of literature to guide our research objectives.

3 Literature Review

We present the findings from the literature review organized around five main topics.

3.1 Potential High-Level Benefits

The private sector has been leading big data research and applications in response to the increase in the amount of data generated online characterized by the three V's: variety, velocity, and volume [7]. Recently, having recognized the value of big data and its potential benefit to society, many governments around the world have invested large sums of money to aggregate and open their datasets, encourage inter-organization cooperation, and establish data analytics applications and frameworks [14].

While parallels can be drawn between the private and public sectors, the objectives of big data implementation in both sectors are different. Governments invest in big data to enable better informed policy decisions and address citizen needs [14], or, in broader terms, to transform an electronic government into a transformational government [15].

Pencheva et al. [9] also classify a body of literature under high-level benefits addressing the effectiveness, efficiency, and the legitimacy that big data applications bring to government initiatives. In particular, it has the potential to invite more participation by citizens and transform governance and decision-making.

Malomo et al. [16], provide considerable evidence on how big data exploitation can generate tangible benefits. The paper resorts to a case study linking transactions relative to students, from different government databases, provided intelligence on their needs. Several other examples of successful implementations and use of data are cited in the literature. For example, the US special force was able to find suspects in the terrorist attack at the Boston Marathon [17]. China has effectively implemented big data centers to analyze the characteristics of crime. In a different initiative, cell phone data records, analyzed in conjunction with traffic safety cameras, were able to profile drivers and identify potential violators of traffic rules.

It is believed big data has the potential to be a "new landmark" in a country's strength as an added dimension to its land, air, sea, and outer space sovereignty [18]. Big data research seems to be the next method of solving a county's toughest challenges. It helps gauge better perceptions of the present to make better decision by making possible the inspection of the granular details of a problem [19].

3.2 Data Quality

Data quality is another prevailing topic, given that big data is often the result of data exhausts – data is generated for logging and other purposes rather than being carefully constructed to guide policy-making [19]. Furthermore, since most big data is generated online, the issues of digital divide becomes prominent, i.e. this representation of samples is biased towards the citizens who are more active users of the Internet.

As a result, it is suggested that a combination of big data, open data, and linked data (BOLD) is the most effective for guiding a process of evidence-based policymaking [20, 21]. In [19], the authors note that BOLD is a better motivator of policy-decisions than citizen surveys, since it avoids the biases of optimistic and pessimistic participants, providing a more objective representation of reality.

There are, however, several issues that lead to misinterpretation of data. For instance, language can be vague, and algorithms may fail at understanding the sentiment as a result. Other issues concerning language involve the noise [19], such as spam, erroneous postings, and irrelevant comments. Moreover, when data from crowdsourcing is utilized it sometimes lack verifiability - it fails to present the context or the metadata which often limits its usability and includes bias as aforementioned [19].

Big data is, by definition, of high variety, and it can take a variety of forms. While this has the potential of enriching the insights garnered, it has the side effect of making models complex and hard to interpret. As such it is important that experiments are performed in a scientific approach, to avoid data misuse, and data omissions in favor of data which supports a favorable outcome. As such, it should be used with established and authoritative data sources to verify patterns and legitimize results [22].

3.3 Privacy Implications

Privacy is a main consideration whenever data pertaining to citizens is collected. This is of particular concern when the data accumulator is a government. There is a big worry about government shaping political opinions or setting agendas by controlling actions of citizens [23].

Actually, there is a fine line between socially beneficial uses of big data and the potential harms to privacy [24]. While laws exist in many countries to anonymize citizen data, the advent of big data has enabled reidentifying small and open data based precisely on it [25]. Shamsi et al. [26] describes 4 types of privacy violations in big data systems based on a literature survey:

1. Tracking by government: surveillance programs run by government that acquire and collect data on individuals from different sources.
2. Data from service providers: non-government organizations and private entities can utilize published government open data to collect information on individuals.
3. Re-identification attacks: correlation of different datasets can identify individuals.
4. Data breaches: security breaches can cause a leak of data.

To counteract these, laws must be drafted to ensure the privacy of individuals. However, in several instances under the pretense of national security, governments

have breached these laws and violated the private lives of individuals under the cover of law. As such, it has been advocated that the laws of privacy should be establish by design rather than by choice. Technological constructions exist to anonymize data and cyber security policies should avoid, or at least limit, data breaches [26].

Conversely, it has been observed that governments are often able to purchase this information from private data collectors, such as Facebook [27]. Indeed, what information a government accumulates is already trumped by the information collected by various private corporations; fear of a big brother is not limited to governments [27].

The literature also distinguishes behavioral big data from inanimate big data [28]. Behavioral big data is data collected on human subjects and their interaction in unprecedented levels of details, and as such it has the potential of causing harm against the subject. The ease of conducting studies at a large scale involving human subjects raises many ethical and even normative questions that researchers are seldom aware of.

3.4 Policy Challenges

Technology has been utilized extensively in the delivery of public policies, namely in the form of e-government initiatives. However, it has had little success, despite of technological advances, in policy design. In [24], the authors describe ways in which big data applications can be used in each step of the policy design cycle based on their claims that it is not feasible to post-evaluate the policy using big data, rather the process of policy making needs to be transformed [24].

Within this context, Janssen et al. [2] identify four types of policy innovations based on the level of involvement of the public: co-creation-based innovation; crowdsourcing-based innovation; service innovation; and policymaking innovation. Co-creation is where data and actors outside of government evaluate government policies and has the least level of involvement, whereas the highest level of involvement policymaking innovation which is analogous to the model suggested in the previous paragraph [29].

Human resources is considered a key challenge in implementing big data based applications for government. Actually, governments often lack expertise, a handicap that can be addressed in two different ways: (1) training and hiring new staff; (2) establishing public/private partnerships. Any partnership with the private sector requires guidelines to expose private entities to big data. This is especially the case where the technical capacities are not available in the government. Indeed, big data based applications require immense data processing power and many governments lack sufficient investment in this arena.

3.5 Big Data in Government

The potential big data-driven applications for the design of public policies is enormous. Some authors, like Hotchl et al. [24] even suggest a revision of the policy cycle such that politicians approach the process of policy formation in a completely different way, entirely driven by data evidence.

Géczy et al. [30] which addresses the problem from various perspective, taking into account the data/technological, process, purpose, and economic dimensions [24].

In [31], Klievink et al. propose the big data use process as a set of activities to handling and making use of big data. In each step of the big data use process, the authors propose a list of activities that are categorizations of data activities contributed by scholars. The process is iterative: output from the last step feeds back into the first one (Fig. 2).

Fig. 2. Big data use process suggested by Klievink *et al.*

4 A Framework Adapted for e-Oman

4.1 Introduction

In light of the topics identified in the literature review, and the challenges of implementing big data in policymaking, we propose a framework based on the authors involvement with e-Oman. The proposed framework attempts to adapt the big data use process of Klievink et al. [31], keeping in mind, however, that this fails to address various data quality challenges identified in the literature review. As such, we propose a 4-step process as follows: *research design*, *data design*, *data analysis*, and *feedback loop*. Here, we take a look at the problem from a slightly higher level to comprehensively encompass the challenges identified in the literature.

While this framework potentially applies agnostically of the governmental or administrative context, we consider it within the context of e-Oman. As mentioned in the Introduction, our objective is to lay the foundations for the implementation of this framework within e-Oman, and as such we consider the current organizational layout and technology foundation existent within the government as an input to our framework design. In the next section, the framework is validated by two key, high-level stakeholders from agency holding the e-Oman Integration platform.

In the next four subsections, we consider each of the proposed steps in detail.

4.2 Research Design

The first step in developing a data-driven application for the governmental domain is its research design. This encompasses study design, ethical conduct and research methods [28]. The possible types of studies vary from prediction of social phenomena, to facilitating information exchange, or even promoting transparency and accountability [19].

A proper construction of the problem for which the evidence is required is thus a necessary first step. It is expected this step is performed in conjunction with the government entity who has requested the evidence. It involves their subject matter experts as well as statistical experts who understand the peculiarities of survey design. Fallacies may arise if the scenario is not properly designed, such as possible confusion between causation and correlation. An often-cited example is estimating the flu epidemic in which Google's algorithm incorrectly estimated the flu based on user searches from the tracked regions [1].

Furthermore, the step encompasses understanding the limits of what can be inferred. This involves testing assumptions and prediction of the models [20], including manual sampling to validate assumptions. It is most important to clearly establish the nature of the outcome of the study in terms of [28]:

- Causality: a relation between a certain event as a cause of a different event.
- Explanation: an understanding of why a given event occurred.
- Prediction: The usage of data to anticipate the occurrence of a given event.
- Correlation: Establishing a link between two events that share a similar stimulus.

4.3 Data Design

Assuring data quality is a major task within this step, through dataset curation which involves data cleansing, validation and anonymization. As mentioned above, big data is often a product of data exhausts [3], i.e. it was not specifically generated for the purpose of being analyzed and has to undergo processing and linkage with other data to become useful.

Equally important is getting access to the data. The datasets used will be a mix of open data, closed data, and social data. While open data is a given, social data often requires special subscriptions. However, closed data is the major obstacle. Policies of different organizations and government agencies have to be respected in terms of data sharing [32]. Request letters have to be sent and links established wherever data is not connected. A major challenge also involves knowing what data is available (Fig. 3).

Fig. 3. Data design steps.

This step also involves data curation, a process along which data can be de-identified, summarized, etc. to maintain privacy and to prepare the data to be analyzed. Models also exist to establish the value of data sets, such as MELODA [33]. Some algorithms for inducing relationships between different datasets include: A/B testing, association rule learning, cluster analysis, neural network analysis, and visualization [3].

4.4 Data Analysis

There are many algorithms that can be implemented to analyze the data. The decision regarding the types of algorithms and data transformations to be used largely depends on the research design and data available. It is out of scope of this paper to survey such algorithms. However, the literature describes and categorizes many of the key algorithms potentially useful within a government context.

A key challenge that arises is the expertise within government organizations to perform this step. Often private public partnerships are needed for this purpose. Those should be carefully prepared and monitored. Another challenge is the storage and processing capabilities required for data analytics.

4.5 Feedback Loop

A feedback loop is a way into which the outcomes of a process are used as input to government policies [21]. Although, in some contexts, this may be considered a source of weakness of evidence-based decision-making [34], a proper feedback loop has the potential of shortening the policy design cycle.

A more accurate definition of a feedback loop is as a way to extend the reuse of the newly found data as an input to the process in a cyclical manner [31]. The two definitions are similar, but the first one cannot always fit the second if the deliverable is concrete.

We opt for the first definition such that a feedback loop is a way of presenting the findings to the policymakers or embedding them in the policy cycle whether or not such findings can be reused as input to a new iteration of the data analytics process or the proposed framework. The feedback loop can be in the form of, dashboards, data, and policy briefs/recommendations.

Dashboards [35] summarize the information that enables policy makers to make informed decisions, and when open to the public, a portal for citizens to scrutinize the government. While dashboards are a way to present data, the data can be summarized in other ways, and can be used as an input in iterations to a continuous monitoring of the policy cycle [31]. Alterations to the policy can be monitored and the data generated used to assess the alterations. A successful use case is president Obama's reelection where the team monitored key indicators and changes to the polls based on the candidate's press releases, campaign visits, and news coverage [19]. Finally, policy briefs would be most effective when external governmental or civil organizations advise policymakers based on the outcome of their application.

Figure 4 compares the steps of our approach against the approach presented by Klievink et al. [31]. As the figure illustrates, the process we propose is more comprehensive in its coverage of the process.

Nonetheless, similar to the earlier process, the proposed framework for governmental data-driven applications design is a general process by which we intend to design our ongoing research in policymaking with respect to e-Oman. The framework is intentionally abstract as the topic addressed is emergent, and the use cases cited in the literature are limited.

While the framework provides an overall process for supporting policymaking within a government context, our survey of literature suggests the work involved in developing an implementation may vary significantly in terms of type and quantity depending on the problem on hand.

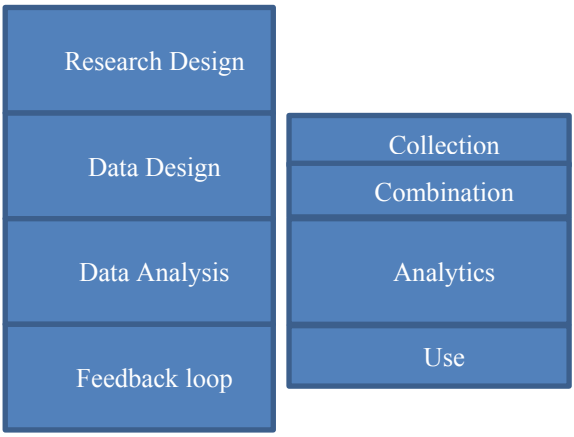

Fig. 4. A comparison between our framework and big data use process.

5 Preliminary Validation

5.1 Stakeholder Interviews

The proposed framework was validated through remote interview with two key stakeholders from the Information Technology Authority (ITA). They are referred in the sequel as (A), director of the e-Oman Integration Platform, and (B), the director general of Infrastructure.

The first question was rather open and intended to validate the very purpose of this work: "(1) What is the potential of the datasets linked to the ITA Integration platform, combined with open data, and social data, be used as a vehicle to verify policy and improve government decision-making?" Both stakeholders (A) and (B) were rather positive, recognizing that the government decision making process will improve due to the consolidation and proper arrangement of data. (A) exemplifies: "If there are new 1000 jobs offered in Muscat, the Ministry of Education can check how many kids they have to increase the number of classes, books, and teachers required for them. The Ministry of Health will be able to prepare its health centers to accommodate the new comers with more doctors, medicines etc." And, in broader terms, "Oman will no longer require to conduct census manually because all required data will be pulled from the already existing registered data in concerned government organizations." Stakeholder (B), on the other hand, recognizes that "there is no plan at this time to expose the data exchanged using integration platform of ITA as these data are property of other organization units". He goes on reinforcing the idea that open and combined data can be a fundamental asset "to take proper strategic decisions".

The second question addressed a major topic related to the item data design in the framework, inquiring about which laws and regulations, if any, need to be drafted to improve government policy-making based on intelligence gathered from data. Gaps on the roles and responsibilities on how to manage data in the e-Oman platform were acknowledge. (B), in particular, emphasized that "data privacy policy need to be clear",

while making clear that some effort will be needed so that "government entities understand the difference between data and information, and start publishing raw data using appropriate channels". Both (A) and (B) were confident that the country will be able to meet this challenge: (A) shared that "if there are any required regulations they will be introduced to accommodate the required changes", while (B) emphasized the need to accompany new regulations with "awareness programs to the public" to foster participation in legislative design based on data evidence.

Inquired about the possible obstacles to intelligent decision-making processes based on data analytics in the Sultanate, (A) stresses the need to clarify the "incremental character" of the framework application. (B), on the other hand, called attention to two main issues, both again related to data: privacy, again, and data ownership ("it is sometimes confusing which government entities should own which data").

We went deeper on the possible constraints, asking to what extent ITA and its government partners have the necessary technical and human resources to design and implement data-driven applications. As anticipated in the framework, both stakeholders recognized the need to involve other partnerships: "We are government, not research centers. So ITA and government should rely on local companies and universities to build the analytical techniques. Another option is to buy (…) but [this] depends on local companies" (A). And "a lot of investment needs to take place to have these skills available" (B).

Finally, we inquired on the main challenges in the application of the framework to concrete projects based on e-Oman. Several factors were mentioned: overcoming the "non-ending bureaucracy" (A), data storage and location ("our main challenge will be having all government data in one location", A); and the need to raise awareness in the public ("it will be sorted out if use cases are brought to see the value of such information produced from open data. B). Security and the need for proper regulations were also mentioned. It was also suggested that the feedback stage in the framework was supported by clear auditing procedures ("a strong and thorough auditing should be implemented to insure confidentiality and integrity of the data", A). AQ3

5.2 Discussion

The stakeholders helped us to validate many of the ideas weaved into the proposed framework as discussed in the previous section.

Issues such as data storage and location, security and regulation were all identified as prerequisites. This has the potential of limiting the types of research questions that can be put forward. Drafting laws and regulations to support such an endeavor is a lengthy activity within a government context.

This increases the envisaged complexity to design the appropriate case study as planned for the future work. In particular, it may require the adoption of the project from a data-owning government entity to analyze the data exchange from a provider level as opposed to an intermediary. The problem will as such be identified in the context of the sponsor for the purpose of a case study.

6 Conclusions and Future Work

In this paper, we lay the groundwork for an application in support of policymaking within government. Our proposed framework for governmental data driven-design is a process based on the considerations put forward by experts in the field, taken within the perspective of government and analyzed in the context of the e-Oman.

The analytical validation establishes the foundation for an ongoing research based on the e-Oman integration platform as a tool to provide and facilitate the small data for policymaking within the Sultanate.

While the framework provides an overall process for supporting policymaking within a government context, our survey of literature and the challenges suggest the work involved in the implementation may vary significantly in terms of type and quantity depending on the problem on hand.

Based on the authors' association with the e-Oman project, an immediate future work is to implement a case study based on the proposed framework in the context of e-Oman and using the e-Oman Integration platform as a leading data source.

We intend to evaluate and, if needed, revise the proposed framework and report our findings to the academic community.

Another area of future work is to test our theories in other policy and social problems based on open data, big data and small data to establish a repository of case studies for the reference of interested researchers. Such a repository will help put forth a more concrete framework and tools that can generalize the common steps.

Acknowledgements. This paper is a result of the project "SmartEGOV: Harnessing EGOV for Smart Governance (Foundations, Methods, Tools)/NORTE-01-0145-FEDER-000037", supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (EFDR).

References

1. Lazer, D., Kennedy, R., King, G., Vespignani, A.: The parable of Google Flu: traps in big data analysis. Science **343**, 1203–1205 (2014)
2. Janssen, M., Konopnicki, D., Snowdon, J., Ojo, A.: Driving public sector innovation using big and open linked data (BOLD). Inf. Syst. Front. **19**, 189–195 (2017)
3. Mergel, I., Rethemeyer, R., Isett, K.: Big data in public affairs. Public Adm. Rev. **76**, 928–937 (2016)
4. Ruhlandt, R.: The governance of smart cities: a systematic literature review. Cities **81**, 1–23 (2018)
5. Elsevier. https://www.elsevier.com/solutions/scopus. Accessed 29 Nov 2018
6. Chadegani, A.A., et al.: A comparison between two main academic literature collections: web of science and scopus databases. Asian Soc. Sci. **9**, 18–26 (2013)
7. Kim, G.-H., Trimi, S., Chung, J.-H.: Big-data applications in the government sector. Commun. ACM **57**, 78–85 (2014)
8. Kummitha, R., Crutzen, N.: How do we understand smart cities? An evolutionary perspective. Cities **67**, 43–52 (2017)

9. Pencheva, I., Esteve, M., Mikhaylov, S.: Big data and AI–a transformational shift for government: so, what next for research? Public Policy Adm. (2018). 0952076718780537 AQ4
10. Webster, J., Watson, R.: Analyzing the past to prepare for the future: writing a literature review. MIS Q. **26**, xiii–xxiii (2002)
11. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
12. Řehůřek, R.: gensim. https://radimrehurek.com/gensim/. Accessed 12 Dec 2018
13. Doc2Vec Tutorial on the Lee Dataset. https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-lee.ipynb. Accessed 12 Dec 2018
14. Gamage, P.: New development: leveraging 'big data' analytics in the public sector. Public Money Manag. **36**, 385–390 (2016)
15. Joseph, R., Johnson, N.: Big data and transformational government. IT Prof. **15**, 43–48 (2013)
16. Malomo, F., Sena, V.: Data intelligence for local government? Assessing the benefits and barriers to use of big data in the public sector. Policy Internet **9**, 7–27 (2017)
17. Kosorukov, A.: Digital government model: theory and practice of modern public administration. J. Legal Ethical Regul. Issues **20** (2017) AQ5
18. Jin, X., Wah, B., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Res. **2**, 59–64 (2015)
19. Taylor, L., Cowls, J., Schroeder, R., Meyer, E.: Big data and positive change in the developing world. Policy Internet **6**, 418–444 (2014)
20. Janssen, M., Kuk, G.: Big and open linked data (BOLD) in research, policy, and practice. J. Organ. Comput. Electron. Commer. **26**, 3–13 (2016)
21. Clarke, A., Margetts, H.: Governments and citizens getting to know each other? Open, closed, and big data in public management reform. Policy Internet **6**, 393–417 (2014)
22. Washington, A.: Government information policy in the era of big data. Rev. Policy Res. **31**, 319–325 (2014)
23. Power, D.: "Big Brother" can watch us. J. Decis. Syst. **25**, 578–588 (2016)
24. Höchtl, J., Parycek, P., Schöllhammer, R.: Big data in the policy cycle: policy decision making in the digital era. J. Organ. Comput. Electron. Commer. **26**, 147–169 (2016)
25. Hardy, K., Maurushat, A.: Opening up government data for big data analysis and public benefit. Comput. Law Secur. Rev. **33**, 30–37 (2017)
26. Shamsi, J., Khojaye, M.: Understanding privacy violations in big data systems. IT Prof. **20**, 73–81 (2018)
27. Lesk, M.: Big data, big brother, big money. IEEE Secur. Priv. **11**, 85–89 (2013)
28. Shmueli, G.: Research dilemmas with behavioral big data. Big Data **5**, 98–119 (2017)
29. Giest, S.: Big data for policymaking: fad or fasttrack? Policy Sci. **50**, 367–382 (2017)
30. Géczy, P.: Big data characteristics. Macrotheme Rev. **3**, 94–104 (2014)
31. Klievink, B., Romijn, B.-J., Cunningham, S., Bruijn, H.: Big data in the public sector: uncertainties and readiness. Inf. Syst. Front. **19**, 267–283 (2017)
32. Bertot, J., Gorham, U., Jaeger, P., Sarin, L., Choi, H.: Big data, open government and e-government: issues, policies and recommendations. Inf. Polity **19**, 5–16 (2014)
33. Abella, A., Ortiz-De-Urbina-Criado, M., De-Pablos-Heredero, C.: A model for the analysis of data-driven innovation and value generation in smart cities' ecosystems. Cities **64**, 47–53 (2017)
34. Poel, M., Meyer, E., Schroeder, R.: Big data for policymaking: great expectations, but with limited progress? Policy Internet **10**, 347–367 (2018)
35. Matheus, R., Janssen, M., Maheshwari, D.: Data science empowering the public: data-driven dashboards for transparent and accountable decision-making in smart cities. Gov. Inf. Q. (2018)

Author Query Form

Book ID : **493546_1_En**

Chapter No : **8**

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,

During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Please check and confirm if the inserted citations of Figures 1–3 are correct. If not, please suggest an alternate citations.	
AQ3	The opening quotes does not have a corresponding closing quotes in the sentence "Several factors were mentioned …". Please insert the quotes in the appropriate position.	
AQ4	Please supply the volume number and page range for References [9, 35].	
AQ5	Please supply the page range for Reference [17].	